Refining the lower bound on the positive eigenvalues of saddle point matrices with insights on the interactions between the blocks

Ruiz, Daniel; Sartenaer, Annick; Tannier, Charlotte

Published in:
SIAM Journal on Matrix Analysis and Applications

DOI:
10.1137/16M108152X

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

SIAM J. MATRIX ANAL. APPL.
Vol. 39, No. 2, pp. 712–736

REFINING THE LOWER BOUND ON THE POSITIVE EIGENVALUES OF SADDLE POINT MATRICES WITH INSIGHTS ON THE INTERACTIONS BETWEEN THE BLOCKS*

DANIEL RUIZ†, ANNICK SARTENAER‡, AND CHARLOTTE TANNIER‡

Abstract. Efficiently solving saddle point systems like Karush–Kuhn–Tucker (KKT) systems is crucial for many algorithms in constrained nonlinear continuous optimization. Such systems can be very ill conditioned, in particular when the (1,1) block has few very small eigenvalues (see Rusten and Winther [*SIAM J. Matrix Anal. Appl.*, 13 (1992), pp. 887–904]). However, it is commonly observed that despite these small eigenvalues, some sort of interaction between this (1,1) block and the (1,2) block actually occurs that may influence strongly the convergence of Krylov subspace methods like MINRES. In this paper, we highlight some aspects of this interaction. We illustrate in particular, with some examples, how and in which circumstances the convergence of MINRES might be affected by these few very small eigenvalues in the (1,1) block. We further derive theoretically a tighter lower bound on the positive eigenvalues of saddle point matrices of the KKT form.

Key words. saddle point systems, ill-conditioning, spectral analysis, minimum residual methods

AMS subject classifications. 15A42, 65F10

DOI. 10.1137/16M108152X

1. Introduction.

We consider the (possibly large and sparse) saddle point linear system

$$\mathcal{A}x = b \equiv \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, with $n \geq m$. We assume that A is symmetric and positive definite and that B has full column rank. These assumptions imply the nonsingularity of $\mathcal{A}$. Such kinds of systems typically arise in constrained nonlinear optimization, as the result of first-order optimality conditions (see [7, section 16.1]), where $\mathcal{A}$ is known as the Karush–Kuhn–Tucker (KKT) matrix. The assumption of positive definiteness of A is met, in particular, when solving strictly convex quadratic optimization problems (see [5] for less restrictive assumptions on A in a constrained optimization context, as well as [1] for a nice survey about saddle point theory and applications). On the application side, systems structured as (1.1) where A is naturally symmetric and positive definite arise in CFD or in magnetostatics, for instance, from the numerical solution of PDEs (see [2, section 5.5] and [9], respectively) or in PDE-constrained optimal control (see [11] and the references therein).

A fundamental result from [10, Lemma 2.1] states that if $\{\mu_i\}_{i=1}^n$ denote the eigenvalues of the symmetric positive definite matrix A and $\{\sigma_i\}_{i=1}^m$ denote the singular

*Received by the editors June 24, 2016; accepted for publication (in revised form) by V. Simoncini February 20, 2018; published electronically April 26, 2018.

http://www.siam.org/journals/simax/39-2/M108152.html

Funding: This work was partially supported by the ANR-BARESAFE project, ANR-11-MONU-004, Programme Modèles Numériques 2011, and the French National Agency for Research. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. The scientific responsibility rests with the authors.

†INPT - ENSEEIHT - IRIT, Toulouse 31071, France (daniel.ruiz@enseeiht.fr).

‡Namur Center for Complex Systems (naXys), University of Namur, Namur 5000, Belgium (annick.sartenaer@unamur.be, charlotte.tannier@unamur.be).

values of the full rank constraint matrix B, then the eigenvalues of $\mathcal{A}$ are bounded within $I^- \cup I^+$, where

$$I^- = \left[\frac{\mu_{\min} - \sqrt{\mu_{\min}^2 + 4\sigma_{\max}^2}}{2}, \frac{\mu_{\max} - \sqrt{\mu_{\max}^2 + 4\sigma_{\min}^2}}{2} \right] \tag{1.2}$$

and

$$I^+ = \left[\mu_{\min}, \frac{\mu_{\max} + \sqrt{\mu_{\max}^2 + 4\sigma_{\max}^2}}{2} \right]. \tag{1.3}$$

As pointed out in [10], the bounds given in (1.2) and (1.3) are sharp, in the sense that there are examples where they are obtained. However, these are worst cases that are not often met in practice, and the purpose of this work is to investigate this in more detail.

Assuming now that the matrix B in (1.1) has orthonormal columns, i.e., $B^T B = I_m$, and that the matrix A has been scaled so as to ensure that its largest eigenvalue $\mu_{\max}$ is close to one, (1.2) and (1.3) then yield the following intervals for the eigenvalues of $\mathcal{A}$ (with $\mu_{\max} = 1$):

$$I^- = \left[\frac{\mu_{\min} - \sqrt{\mu_{\min}^2 + 4}}{2}, \frac{1 - \sqrt{5}}{2} \right] \quad \text{and} \quad I^+ = \left[\mu_{\min}, \frac{1 + \sqrt{5}}{2} \right]. \tag{1.4}$$

That is, by orthonormalizing the columns of B and scaling the matrix A, the negative eigenvalues of $\mathcal{A}$ (in I^-) are guaranteed to be well bounded and away from zero, and the positive ones (in I^+) are guaranteed to be well bounded too. This, however, does not exclude the possibility for $\mathcal{A}$ to have very small positive eigenvalues if the smallest eigenvalue of A, $\mu_{\min}$, is very close to zero.

Let us illustrate this possibility on the 3×3 matrix

$$\mathcal{A}_{cs} = \begin{pmatrix} \mu & 0 & c \\ 0 & 1 & s \\ c & s & 0 \end{pmatrix}, \tag{1.5}$$

where $0 < \mu < 1$ and $c^2 + s^2 = 1$. First, observe that if $c^2 = 0$, then the eigenvalues of $\mathcal{A}_{cs}$ are $\{(1 - \sqrt{5})/2, \mu, (1 + \sqrt{5})/2\}$, and its smallest positive eigenvalue is directly given by the smallest positive eigenvalue in its (1,1) block (this illustrates the worst cases mentioned above). Now, if we take $c^2 = 1$, then the eigenvalues of $\mathcal{A}_{cs}$ are $\{(\mu - \sqrt{\mu^2 + 4})/2, 1, (\mu + \sqrt{\mu^2 + 4})/2\}$ and thus well bounded and isolated away from zero no matter how close μ is to zero. A simple analytical analysis, given in Appendix A, also shows that the smallest positive eigenvalue of $\mathcal{A}_{cs}$ is $\mathcal{O}\left(c^2/(1 + c^2)\right)$, and thus bounded away from zero, when $\mu \ll c^2 \leq 1$. This shows the role played by the constraint block $B = (c\ s)^T$ (even orthonormalized) in relaying the bad conditioning (if μ is close to zero) of the (1,1) block into the saddle point matrix $\mathcal{A}_{cs}$ and highlights the existence of some sort of interaction between the blocks in $\mathcal{A}_{cs}$. Note that $c \in [0, 1]$, whose size plays a fundamental role in this interaction, is nothing else than the cosine of the principal angle between $\mathcal{I}m(B)$ and the invariant subspace associated to the smallest eigenvalue μ of the (1,1) block A (see [4, section 6.4.3]).

The above example motivates the purpose of this paper, which is to present some insights on the interaction between the blocks in saddle point systems of the form

(1.1) and to identify some circumstances in which the smallest eigenvalues contained in the (1,1) block will, or will not, spoil the convergence of Krylov subspace methods like MINRES. It is indeed well known that, for a saddle point matrix $\mathcal{A}$, the effective condition number of its (1,1) block A on the null space of B^T plays a central role. Indeed, even if A is semipositive definite with several zero eigenvalues, the matrix $\mathcal{A}$ can be well conditioned and easy to work with, as long as the null spaces of A and B^T are well separated. With respect to the above discussion, this corresponds to cosines close to 1 between the range of B and the null space of A. These considerations will lead us to refine the lower bound in (1.3) given in [10] on the positive eigenvalues of saddle point matrices of the form (1.1).

The paper is organized as follows. Section 2 illustrates in two different ways, and through the convergence of MINRES, the influence of the cosines of the principal angles between $\mathcal{I}m(B)$ and the invariant subspace associated to the smallest eigenvalues of the (1,1) block A. In section 3, we present theoretical results refining the lower bound $\mu_{\min}$ of the right interval in (1.4). We shall restrict our analysis to the case where the (1,1) block largest eigenvalue is scaled to 1 and the constraint matrix has orthonormal columns. These assumptions can be met in practice when preconditioning the saddle point matrix $\mathcal{A}$ (see section 2.2). Before getting into the details of the theroretical analysis itself, which are conducted throughout subsections 3.1–3.3, we suggest that the reader gives first a quick look at subsection 3.4, in which the conclusions are raised and a plot of the different important steps within the analysis is recalled, so as to clearly have in mind the target and the reasoning as well. Finally, in section 4, we conclude with our main result and some prospective remarks.

2. Interaction between the blocks in KKT matrices. In the introduction, we have raised the possible influence of the cosines of the principal angles between $\mathcal{I}m(B)$ and the invariant subspace associated to the smallest eigenvalues of the (1,1) block A and shown that they can have a sizeable impact on the lower bound of the positive interval in (1.4) for the spectrum of the KKT matrix $\mathcal{A}$ in (1.1). We now present some illustrations of this effect through the convergence of MINRES.

2.1. Illustration on a toy example. We first introduce a "*hand-made*" example to illustrate how these cosines values can impact and eventually spoil the convergence of MINRES. To build the matrix $\mathcal{A}$, we consider a diagonal matrix A of order $n = 500$ with diagonal entries in $]0, 1]$ and such that

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix},$$

with $A_1 \in \mathbb{R}^{5\times 5}$ being set as

$$\begin{aligned} A_1 &= \operatorname{diag}(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) \\ &= \operatorname{diag}(10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 10^{-1}) \end{aligned}$$

and $A_2 \in \mathbb{R}^{(n-5)\times(n-5)}$ being a diagonal matrix with uniform values from the interval $[a, b] = [0.101, 1]$ and randomly generated by the MATLAB code

```
diag(a + (b-a).*rand(n-5,1)).
```

The eigenvectors of A are therefore given by the canonical basis of $\mathbb{R}^n$. The matrix $B \in \mathbb{R}^{n\times m}$, where $m = 200$, is set to

$$B = \begin{bmatrix} C & 0 \\ B_1 S & B_2 \end{bmatrix},$$

TABLE 2.1
Values of the cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_1)$ for four configurations.

	(a)	(b)	(c)	(d)
$\cos\theta_1$	0.3	0.3	10^{-4}	0
$\cos\theta_2$	0.3	10^{-3}	10^{-3}	0
$\cos\theta_3$	0.3	0.3	10^{-2}	0
$\cos\theta_4$	0.3	0.3	10^{-1}	0
$\cos\theta_5$	0.3	0.3	0.3	0

where $C = \mathrm{diag}\{\cos\theta_i\}_{i=1}^5$, $S = \mathrm{diag}\{\sin\theta_i\}_{i=1}^5$, $B_1 \in \mathbb{R}^{(n-5)\times 5}$, $B_2 \in \mathbb{R}^{(n-5)\times(m-5)}$, with $Q = [B_1\ B_2] \in \mathbb{R}^{(n-5)\times m}$ dense and satisfying $Q^TQ = I_m$ so as to ensure that B has orthonormal columns. We also have, considering $U_1 = \begin{bmatrix} I_5 & 0 \end{bmatrix}^T$ the set of eigenvectors corresponding to the first five eigenvalues of A,

$$B^TU_1 = \begin{bmatrix} C \\ 0 \end{bmatrix},$$

so that we explicitly get—and this is done on purpose in this particular example—a one-to-one matching between eigenvectors versus principal vectors and eigenvalues versus cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_1)$. We consider four different configurations (a), (b), (c), and (d) for

$$C = \mathrm{diag}\{\cos\theta_i\}_{i=1}^5,$$

with values of the cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_1)$ given in Table 2.1. Figure 2.1 illustrates and compares the impact of the values of the cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_1)$ on the behavior of MINRES applied to this toy KKT matrix for the four cases. For reference, we indicate with the dashed (red) curve on each graph the convergence profile of MINRES when all the five cosines are set to one, and in all cases the iterations are stopped when the scaled residual $\|r^k\|/\|r^0\|$ in 2-norm is less than 10^{-10}. For particular values of the cosines, the phenomenon of plateau occurs and, as indicated by the preliminary comments made in the introduction, we can observe that when the squares of the cosines of some principal angles are equal to the corresponding eigenvalues (in this particular one-to-one cosine-eigenvalue matching test example), the corresponding bad conditioning of A is showing up. For instance, if we change the value of the second cosine $\cos\theta_2$ from 0.3 to 10^{-3} (corresponding to the square root of the corresponding eigenvalue) between situations (a) and (b), the speed of convergence of MINRES is disrupted. One such phenomenon of plateau in the convergence curve occurs in Figure 2.1, case (b). Case (c) corresponds to a generalized case where each of the five values of the cosines reveals the corresponding eigenvalue in A, leading to five phenomena of plateau in Figure 2.1, case (c). What is interesting to observe in case (d) is that setting even the five cosine values to zero, which corresponds to the case where the bound from Rusten and Winther is sharp, does not change the behavior of MINRES much compared to case (c) (in which the squares of the cosines are already at the level of their corresponding eigenvalue). At last, in case (a), the square of theses cosines is of the same order of magnitude as that of the smallest eigenvalue in A_2. We can see that the behavior of MINRES exhibits a linear rate of convergence and does not depart too much from the reference case in which all theses cosines are actually set to one. This reference case

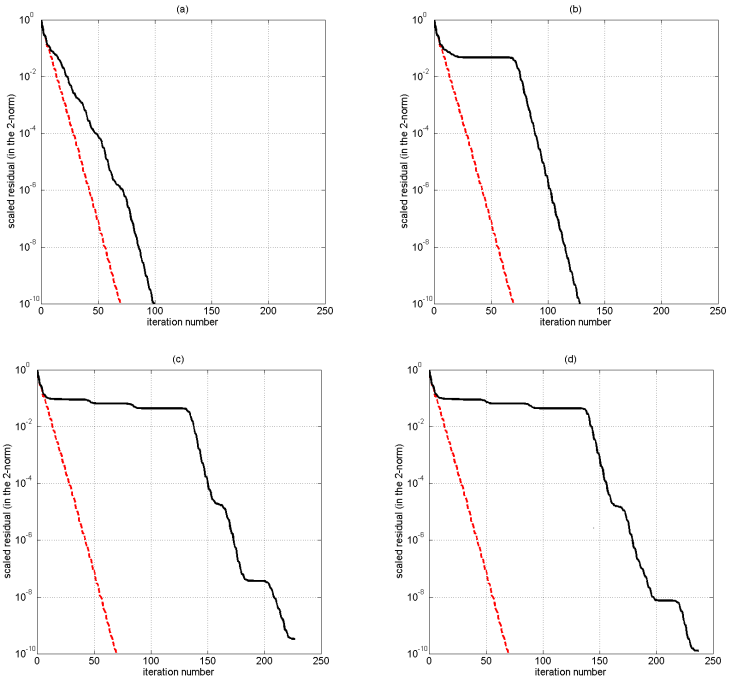

FIG. 2.1. *Convergence profiles (2-norm of scaled residuals) for the four cosines configurations.*

corresponds to the extreme opposite case where the invariant subspace associated to the five smallest eigenvalues is actually orthogonal to $\mathcal{K}er(B^T)$ and thus not active in the degrees of freedom of the constraint set.

2.2. Varying the constraint matrix. The various observations previously raised suggest that if all the cosines of the principal angles between $\mathcal{I}m(B)$ and the invariant subspace associated to the smallest eigenvalues in the (1,1) block A are large enough, the convergence of MINRES preconditioned with the classical block diagonal preconditioner

$$\mathcal{P} = \begin{bmatrix} \mu_{\max} I_n & 0 \\ 0 & \frac{1}{\mu_{\max}} B^T B \end{bmatrix}, \tag{2.1}$$

which essentially scales the largest eigenvalue of A and orthonormalizes the constraint matrix B, should be reasonably fast, independently of any consideration with respect to the ill-conditioning in A.

To illustrate this point, we now consider a KKT matrix built in the following way. We first generate in MATLAB a symmetric positive definite test matrix of order $n = 300$, with a relatively well clustered spectrum showing 42 eigenvalues less than $\gamma = \frac{\mu_{\max}}{100} \approx 3.8\,10^{-2}$ but with extreme eigenvalues $\mu_{\min} = 1.7\,10^{-7}$ and $\mu_{\max} = 3.8$. The condition number of this (1,1) block test matrix A is then $2.2\,10^7$. The constraint matrix $B \in \mathbb{R}^{300\times 150}$ is generated by means of the MATLAB function `sprandn` (with a density of 0.05 and a condition number of 10^4). We then slightly modify this constraint matrix B into $\widetilde{B}$ so as to enforce the cosines of the principal angles between $\mathcal{I}m(\widetilde{B})$ and $\mathcal{I}m(U_\gamma)$, the invariant subspace associated to the $p = 42$

TABLE 2.2
Values of the $\ell = 22$ cosines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$ below 0.362.

$\cos\theta_i$					
$4.98\,10^{-3}$	$6.88\,10^{-3}$	$8.21\,10^{-3}$	$1.23\,10^{-2}$	$1.72\,10^{-2}$	$2.39\,10^{-2}$
$3.03\,10^{-2}$	$4.00\,10^{-2}$	$4.32\,10^{-2}$	$4.62\,10^{-2}$	$5.51\,10^{-2}$	$6.05\,10^{-2}$
$7.20\,10^{-2}$	$7.83\,10^{-2}$	$8.86\,10^{-2}$	$1.46\,10^{-1}$	$1.49\,10^{-1}$	$1.69\,10^{-1}$
$2.02\,10^{-1}$	$2.83\,10^{-1}$	$3.22\,10^{-1}$	$3.33\,10^{-1}$		

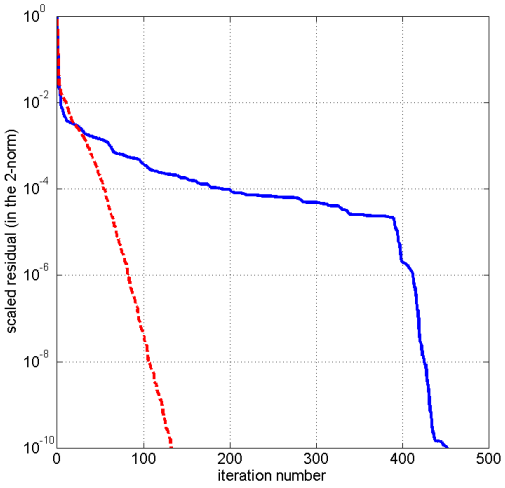

FIG. 2.2. *Convergence profiles of preconditioned* MINRES *(with $\mathcal{P}$). Case with constraint matrix B in solid (blue) curve, and case with constraint matrix $\widetilde{B}$ in dashed (red) curve.*

eigenvalues smaller than $\gamma \approx 3.8\,10^{-2}$, to be larger than a certain threshold, e.g.,

$$\min\{\cos\widetilde{\theta}_i\}_{i=1}^{p} \geq 2\sqrt{\frac{\gamma}{\alpha}} \simeq 0.362\,,$$

with $\alpha = 1.16$ corresponding to the average of the eigenvalues of A above γ.

The details on how the constraint matrix B is modified into $\widetilde{B}$ are given in [12, section 6.2.2]. This results from some particular linear combination between B and U_γ that essentially preserves the normal equations of B, so that $B^TB = \widetilde{B}^T\widetilde{B}$, and ensures the above condition on the cosines.

In the case of this test example, amongst the $p = 42$ cosines, only the $\ell = 22$ smallest ones are actually below the threshold $2\sqrt{\gamma/\alpha} \simeq 0.362$, and these are displayed in increasing order in Table 2.2. In the modified constraint matrix $\widetilde{B}$, all these 22 cosines have been raised to the value 0.362 explicitly, leaving the other principal angles unchanged. Figure 2.2 shows the convergence profiles of MINRES preconditioned with $\mathcal{P}$ in both cases (e.g., with either B or $\widetilde{B}$). It is important to note that the preconditioning matrix (2.1) is the same in both cases, since $B^TB = \widetilde{B}^T\widetilde{B}$ and the (1,1) block is unchanged. In the case with large enough cosines, the smallest eigenvalues in the (1,1) block A have almost no impact, and a simple implicit orthonormalization of the constraints (with $\mathcal{P}$) is enough to reach linear convergence in MINRES.

3. A refined eigenvalue bound for KKT matrices. In this section, we aim at refining the lower bound $\mu_{\min}$ of the positive interval in (1.4) through a theoretical analysis in terms of cosines of principal angles between the subspace spanned by the constraint equations and the subspace spanned by the eigenvectors associated to the smallest eigenvalues of the (1,1) block A. Doing so, we expect to clarify those situations where this lower bound is guaranteed to be bounded away from zero.

Before going into the theoretical developments of the following subsections, we first introduce some notation and the specific assumptions underlying the analysis. In what follows, we consider a particular case of the KKT matrix,

$$\bar{\mathcal{A}} = \begin{bmatrix} A & Q \\ Q^T & 0 \end{bmatrix}, \tag{3.1}$$

with $A \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times m}$ satisfying $Q^TQ = I_m$, and $n \geq m$. This corresponds to a constraint matrix B in (1.1) whose columns are orthonormal. We also assume that some scaling has also been applied so that the largest eigenvalue of the symmetric positive definite matrix A is equal to one ($\mu_{\max} = 1$).

To analyze the spectrum of $\bar{\mathcal{A}}$ in (3.1), we rewrite this matrix into two successively similar matrices. We first consider the eigendecomposition of the (1,1) block

$$A = U\Delta U^T,$$

where the diagonal matrix $\Delta \in \mathbb{R}^{n\times n}$ contains the eigenvalues $\{\mu_i\}_{i=1}^n$ of A and the orthonormal matrix $U \in \mathbb{R}^{n\times n}$ contains the associated orthonormal set of eigenvectors, and observe that

$$\begin{aligned} \begin{bmatrix} A & Q \\ Q^T & 0 \end{bmatrix} &= \begin{bmatrix} U\Delta U^T & Q \\ Q^T & 0 \end{bmatrix} \\ &= \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Delta & U^TQ \\ Q^TU & 0 \end{bmatrix} \begin{bmatrix} U^T & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \tag{3.2}$$

We next split the spectrum of A in two parts, with $\Delta_\gamma \in \mathbb{R}^{p\times p}$ the diagonal matrix containing the p eigenvalues $0 < \mu_{\min} = \mu_1 \leq \cdots \leq \mu_p$ strictly less than a given positive number $\gamma \in [\mu_{\min}, 1]$, and with $\widetilde{\Delta}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$ the diagonal matrix containing all the other $(n-p)$ eigenvalues $\gamma \leq \mu_{p+1} \leq \cdots \leq \mu_n = \mu_{\max} = 1$. The matrix $U = [U_\gamma,\ \widetilde{U}_\gamma] \in \mathbb{R}^{n\times n}$ is orthogonal, where the columns of the rectangular matrices $U_\gamma \in \mathbb{R}^{n\times p}$ and $\widetilde{U}_\gamma \in \mathbb{R}^{n\times(n-p)}$ are the orthonormal sets of eigenvectors corresponding to Δ_γ and $\widetilde{\Delta}_\gamma$, respectively.

We now introduce the cosine-sine (CS) decomposition of matrix $K \in \mathbb{R}^{m\times n}$ defined as

$$K = Q^TU = [Q^TU_\gamma,\ Q^T\widetilde{U}_\gamma] = [K_\gamma,\ \widetilde{K}_\gamma], \tag{3.3}$$

where $Q = B(B^TB)^{-1/2} \in \mathbb{R}^{n\times m}$ satisfies $Q^TQ = I_m$ by definition, and $U = [U_\gamma,\ \widetilde{U}_\gamma]$ is an orthogonal matrix made with those eigenvectors of A. The columns of K^T are orthonormal, implying that $K_\gamma K_\gamma^T + \widetilde{K}_\gamma\widetilde{K}_\gamma^T = I_m$. If we next complete the matrix K^T by $m-n$ orthonormal columns to provide an orthogonal matrix of $\mathbb{R}^{n\times n}$ and if we apply the CS decomposition as in [8, section 4], one can guarantee the existence of orthogonal matrices $V_\gamma \in \mathbb{R}^{p\times p}$, $\widetilde{V}_\gamma \in \mathbb{R}^{(n-p)\times(n-p)}$, and $W \in \mathbb{R}^{m\times m}$ such that

$$V_\gamma^T K_\gamma^T W = \mathcal{C} = \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_r) \in \mathbb{R}^{p\times m}, \quad r = \min\{p, m\}, \tag{3.4}$$

and

$$\widetilde{V}_\gamma^T \widetilde{K}_\gamma^T W = \mathcal{S} = \operatorname{diag}(\sin\theta_1, \ldots, \sin\theta_q) \in \mathbb{R}^{(n-p)\times m}, \quad q = \min\{n-p, m\}, \tag{3.5}$$

where $\mathcal{C}^T\mathcal{C} + \mathcal{S}^T\mathcal{S} = I_m$. The singular values $\cos\theta_i$ and $\sin\theta_i$ of K_γ^T and $\widetilde{K}_\gamma^T$, respectively, are cosines and sines satisfying (without loss of generality)

$$1 \geq \cos\theta_1 \geq \cdots \geq \cos\theta_r \geq 0 \qquad \text{and} \qquad 0 \leq \sin\theta_1 \leq \cdots \leq \sin\theta_q \leq 1.$$

Among these values, $\min\{r, q\}$ correspond to the cosines and sines of the principal angles between $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, the other values being equal to either zero or one, depending on the dimensions p, m, and n. The associated $\min\{r, q\}$ principal vectors (see [4, section 6.4.3]) are defined by the $\min\{r, q\}$ first columns of matrix QW and matrix $U_\gamma V_\gamma$ in $\mathcal{I}m(B)$ and $\mathcal{I}m(U_\gamma)$, respectively.

For simplicity, we shall restrict ourselves to the case where $p < m$ and $m < n - p$, so that $r = p$ and $q = m$, and

$$\mathcal{C} = \begin{bmatrix} C & 0 \end{bmatrix} \in \mathbb{R}^{p\times m} \quad \text{and} \quad \mathcal{S} = \begin{bmatrix} S & 0 \\ 0 & I_{m-p} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n-p)\times m}, \tag{3.6}$$

with $C \in \mathbb{R}^{p\times p}$ and $S \in \mathbb{R}^{p\times p}$. Extracting K_γ and $\widetilde{K}_\gamma$ from (3.4) and (3.5) also yields $K_\gamma = W\mathcal{C}^T V_\gamma^T$ and $\widetilde{K}_\gamma = W\mathcal{S}^T\widetilde{V}_\gamma^T$. Using these expressions and remembering that $Q^T U = K = [K_\gamma,\ \widetilde{K}_\gamma]$ by (3.3), we rewrite the central matrix in (3.2) in terms of cosines and sines as

$$\begin{bmatrix} \Delta_\gamma & 0 & V_\gamma \mathcal{C} W^T \\ 0 & \widetilde{\Delta}_\gamma & \widetilde{V}_\gamma \mathcal{S} W^T \\ W\mathcal{C}^T V_\gamma^T & W\mathcal{S}^T\widetilde{V}_\gamma^T & 0 \end{bmatrix}.$$

We next obtain

$$\begin{bmatrix} V_\gamma^T & 0 & 0 \\ 0 & \widetilde{V}_\gamma^T & 0 \\ 0 & 0 & W^T \end{bmatrix} \begin{bmatrix} \Delta_\gamma & 0 & V_\gamma \mathcal{C} W^T \\ 0 & \widetilde{\Delta}_\gamma & \widetilde{V}_\gamma \mathcal{S} W^T \\ W\mathcal{C}^T V_\gamma^T & W\mathcal{S}^T\widetilde{V}_\gamma^T & 0 \end{bmatrix} \begin{bmatrix} V_\gamma & 0 & 0 \\ 0 & \widetilde{V}_\gamma & 0 \\ 0 & 0 & W \end{bmatrix} = \begin{bmatrix} M_\gamma & 0 & \mathcal{C} \\ 0 & \widetilde{M}_\gamma & \mathcal{S} \\ \mathcal{C}^T & \mathcal{S}^T & 0 \end{bmatrix},$$

where

$$M_\gamma = V_\gamma^T \Delta_\gamma V_\gamma \quad \text{and} \quad \widetilde{M}_\gamma = \widetilde{V}_\gamma^T \widetilde{\Delta}_\gamma \widetilde{V}_\gamma, \tag{3.7}$$

which is also similar to $\bar{\mathcal{A}}$ due to the fact that matrices $V_\gamma, \widetilde{V}_\gamma$, and W are orthogonal. Finally, using (3.6), we end up with the following similar matrix:

$$\left[\begin{array}{c|cc|cc} M_\gamma & \multicolumn{2}{c|}{0} & C & 0 \\ \hline & & & S & 0 \\ 0 & \multicolumn{2}{c|}{\widetilde{M}_\gamma} & 0 & I \\ & & & 0 & 0 \\ \hline C & S & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \end{array}\right]. \tag{3.8}$$

We are now ready to analyze the minimal positive eigenvalue of this matrix in block form. In section 3.1, we first deduce some general spectral relations before focusing, in section 3.2, on the positive eigenvalues of the matrix (3.1) that are smaller than the value of $\gamma/2$. Finally, based on these spectral relations, we successively define in section 3.3 two constrained optimization problems whose minimal value will lead to a refined lower bound for the positive interval in (1.4). The norm considered in the following is the 2-norm $\|\cdot\|_2$, and we use in short the notation $\|\cdot\|$.

3.1. General spectral relations. Let $\lambda \in \mathbb{R}$ denote an eigenvalue of the matrix given in (3.8), with the associated eigenvector $\begin{bmatrix} x_1 & x_2 & y_1 & y_2 \end{bmatrix}^T$ in which $x_1 \in \mathbb{R}^p$, $x_2 \in \mathbb{R}^{n-p}$, $y_1 \in \mathbb{R}^p$, and $y_2 \in \mathbb{R}^{m-p}$. We then have the following equalities:

$$M_\gamma x_1 + C y_1 = \lambda x_1 \quad \text{and} \quad \widetilde{M}_\gamma x_2 + \begin{bmatrix} S y_1 \\ y_2 \\ 0 \end{bmatrix} = \lambda x_2, \tag{3.9}$$

together with

$$C x_1 + \begin{bmatrix} S & 0 & 0 \end{bmatrix} x_2 = \lambda y_1 \quad \text{and} \quad \begin{bmatrix} 0 & I & 0 \end{bmatrix} x_2 = \lambda y_2. \tag{3.10}$$

Let us introduce now the positive quantities $\rho_1 \in [\mu_{\min}, \mu_p]$ and $\rho_2 \in [\mu_{p+1}, 1]$ satisfying

$$x_1^T M_\gamma x_1 = \rho_1 \|x_1\|^2 \quad \text{and} \quad x_2^T \widetilde{M}_\gamma x_2 = \rho_2 \|x_2\|^2. \tag{3.11}$$

Note that if both $x_1 \neq 0$ and $x_2 \neq 0$, then ρ_1 and ρ_2 are Rayleigh quotients and satisfy $\rho_1 \in [\mu_{\min}, \mu_p]$ and $\rho_2 \in [\mu_{p+1}, 1]$ by (3.7). Otherwise, if $x_1 = 0$ or $x_2 = 0$ or both, it is always possible to find positive quantities $\rho_1 \in [\mu_{\min}, \mu_p]$ and $\rho_2 \in [\mu_{p+1}, 1]$ to satisfy (3.11).

The following lemma gives some spectral relations, which will be useful in the next sections (for the proof, see Appendix B).

LEMMA 3.1. *Let $\lambda \in \mathbb{R}$ be an eigenvalue of the matrix* (3.8) *associated to the eigenvector $x = \begin{bmatrix} x_1 & x_2 & y_1 & y_2 \end{bmatrix}^T$ with $x_1 \in \mathbb{R}^p$, $x_2 \in \mathbb{R}^{n-p}$, $y_1 \in \mathbb{R}^p$, and $y_2 \in \mathbb{R}^{m-p}$. Then x satisfies the following relations:*

$$\lambda^2 \|x_1\|^2 = \lambda \rho_1 \|x_1\|^2 + x_1^T C^2 x_1 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2, \tag{3.12}$$

$$\lambda^2 \|x_2\|^2 = \lambda \rho_2 \|x_2\|^2 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2 + x_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{3.13}$$

$$\lambda^2 \|y_1\|^2 = x_1^T C^2 x_1 + 2 x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2 + x_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{3.14}$$

$$\lambda^2 \|y_2\|^2 = x_2^T \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} x_2, \tag{3.15}$$

where $\rho_1 \in [\mu_{\min}, \mu_p]$ and $\rho_2 \in [\mu_{p+1}, 1]$ satisfy (3.11).

3.2. Specific relations for small positive eigenvalues. As we have seen in (1.4), the lower bound of the interval associated to the positive eigenvalues of

$\bar{\mathcal{A}}$ given in (3.1) is not necessarily isolated away from zero (especially when $\mu_{\min}$ is extremely small). In this section, we thus focus our analysis on the eigenvalues of $\bar{\mathcal{A}}$ smaller than the given threshold γ and we deduce, from Lemma 3.1, specific relations associated to these eigenvalues. As we will see, small positive eigenvalues of $\bar{\mathcal{A}}$ require that the weights in the eigencomponents of the associated eigenvectors are more important on the part relative to the small eigenvalues of A. In particular, for the existence of a positive eigenvalue $\bar{\lambda}$ less than $\gamma/2$, we will demonstrate that it is mandatory to have $\|\bar{x}_1\| > \|\bar{x}_2\|$ within the blocks of the corresponding eigenvector $\bar{x} = \begin{bmatrix}\bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2\end{bmatrix}^T$.

We therefore start our analysis by assuming that the eigenvalue problem defined by the equalities in (3.9) and (3.10) has a positive eigenvalue $\bar{\lambda} \geq 0$ such that $\bar{\lambda} < \gamma/2$, and we shall then study, under this hypothesis, possible lower bound values for this eigenvalue $\bar{\lambda}$. We also make the assumption that the minimum cosine value in C is strictly positive; otherwise, we already know, from the example in the introduction, that the positive lower bound from Rusten and Winther can be sharp. The following lemma summarizes the various necessary conditions that must be met due to the existence of such an eigenvalue $\bar{\lambda} < \gamma/2$, together with strictly positive cosines (for the proof, see Appendix C).

LEMMA 3.2. *Assume the matrix* (3.8) *has a positive eigenvalue* $\bar{\lambda}$ *satisfying* $\bar{\lambda} < \gamma/2$, *with* $\gamma \in [\mu_{\min}, 1]$, *and with the associated eigenvector* $\bar{x} = \begin{bmatrix}\bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2\end{bmatrix}^T$, *where* $\bar{x}_1 \in \mathbb{R}^p$, $\bar{x}_2 \in \mathbb{R}^{n-p}$, $\bar{y}_1 \in \mathbb{R}^p$, *and* $\bar{y}_2 \in \mathbb{R}^{m-p}$. *Let also* C *and* $S \in \mathbb{R}^{p\times p}$ *given by* (3.6) *satisfy* $c_{\min} = \min_{i=1:p}\{\cos\theta_i\} > 0$. *Then it is necessary to have* $\|\bar{x}_1\| > \|\bar{x}_2\| > 0$ *and* $c_{\min} < 1$.

These necessary conditions allow us to divide the previous relations by either $\|\bar{x}_1\|$ or $\|\bar{x}_2\|$ and to derive some new relations based only on some specific scalar quantities in prescribed intervals. These scalar quantities actually correspond to energy estimates, in the spirit of what is done by Rusten and Winter to derive their well-known bounds on the eigenvalues of the KKT matrix. We first introduce these specific energy estimates associated to the existing eigenvalue $0 < \bar{\lambda} < \gamma/2$, together with the assumption that $c_{\min} > 0$ (for which $\|\bar{x}_1\| > \|\bar{x}_2\| > 0$, as guaranteed by Lemma 3.2):

$$\bar{\omega} = \frac{\|\bar{x}_1\|^2}{\|\bar{x}_2\|^2} > 1, \tag{3.16}$$

$$\bar{\rho}_1 = \frac{\bar{x}_1^T M_\gamma \bar{x}_1}{\|\bar{x}_1\|^2} \in [\mu_{\min}, \mu_p] \quad \text{and} \quad \bar{\rho}_2 = \frac{\bar{x}_2^T \widetilde{M}_\gamma \bar{x}_2}{\|\bar{x}_2\|^2} \in [\mu_{p+1}, 1], \tag{3.17}$$

together with

$$\bar{\alpha} = \frac{\bar{x}_1^T C^2 \bar{x}_1}{\|\bar{x}_1\|^2} \in [c_{\min}^2, 1] \quad \text{and} \quad \bar{\beta} = \frac{\bar{x}_2^T \begin{bmatrix} S^2 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_2\|^2} \in [0, 1] \tag{3.18}$$

and

$$\bar{\tau} = \frac{\bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2}{\|\bar{x}_1\|^2}. \tag{3.19}$$

In the next theorem, we transform the relations derived in Lemma 3.1 and introduce scalar relations based on the above quantities $\bar{\omega}$, $\bar{\rho}_1$, $\bar{\rho}_2$, $\bar{\alpha}$, $\bar{\beta}$, and $\bar{\tau}$. The key point is actually to identify some particular inequalities, in order to derive a set of nonlinear equations and inequations that we shall study in the next section to obtain a lower bound on the given positive eigenvalue $0 < \bar{\lambda} < \gamma/2$.

THEOREM 3.3. *Assume that the matrix in* (3.8) *has a positive eigenvalue* $\bar{\lambda}$ *satisfying* $\bar{\lambda} < \gamma/2$*, with* $\gamma \in [\mu_{\min}, 1]$*, and with the associated eigenvector*

$$\bar{x} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{y}_1 & \bar{y}_2 \end{bmatrix}^T,$$

with $\bar{x}_1 \in \mathbb{R}^p$*,* $\bar{x}_2 \in \mathbb{R}^{n-p}$*,* $\bar{y}_1 \in \mathbb{R}^p$*, and* $\bar{y}_2 \in \mathbb{R}^{m-p}$*. Let* $\bar{\omega}$*,* $\bar{\rho}_1$*,* $\bar{\rho}_2$*,* $\bar{\alpha}$*,* $\bar{\beta}$*, and* $\bar{\tau}$ *be given by* (3.16)–(3.19)*, respectively, and let* $c_{\min} > 0$*. We then have*

$$-\bar{\lambda}^2 + \bar{\lambda}\bar{\rho}_1 + \bar{\alpha} + \bar{\tau} = 0, \tag{3.20}$$

$$-\bar{\lambda}^2 + \bar{\lambda}\bar{\rho}_2 + \bar{\beta} + \bar{\tau}\bar{\omega} = 0, \tag{3.21}$$

$$\bar{\tau} + \bar{\alpha} > 0, \tag{3.22}$$

$$\bar{\tau}\bar{\omega} + \bar{\beta} < 0, \tag{3.23}$$

$$\bar{\tau}^2\bar{\omega} \leq \bar{\alpha}\bar{\beta}. \tag{3.24}$$

Proof. It is straightforward to derive (3.20) and (3.21) by dividing (3.12) and (3.13) by $\|\bar{x}_1\|^2$ and $\|\bar{x}_2\|^2$, respectively (where $\bar{x}_1 \neq 0$ and $\bar{x}_2 \neq 0$ by Lemma 3.2).

We next prove (3.22). Adding (3.14) to (3.15) gives

$$0 \leq \bar{\lambda}^2(\|\bar{y}_1\|^2 + \|\bar{y}_2\|^2) = \bar{\alpha}\|\bar{x}_1\|^2 + 2\bar{\tau}\|\bar{x}_1\|^2 + \bar{\beta}\|\bar{x}_2\|^2, \tag{3.25}$$

which again, with the quantities above, can be written as

$$(\bar{\alpha} + \bar{\tau})\|\bar{x}_1\|^2 + (\bar{\tau}\bar{\omega} + \bar{\beta})\|\bar{x}_2\|^2 \geq 0. \tag{3.26}$$

Observing also that (3.13) can be written as $(\bar{\tau}\bar{\omega} + \bar{\beta})\|\bar{x}_2\|^2 = \bar{\lambda}(\bar{\lambda} - \bar{\rho}_2)\|\bar{x}_2\|^2$, (3.26) becomes

$$(\bar{\alpha} + \bar{\tau})\|\bar{x}_1\|^2 + \bar{\lambda}(\bar{\lambda} - \bar{\rho}_2)\|\bar{x}_2\|^2 \geq 0,$$

or equivalently, after division by $\|\bar{x}_2\|^2$, $(\bar{\alpha} + \bar{\tau})\bar{\omega} \geq \bar{\lambda}(\bar{\rho}_2 - \bar{\lambda})$. Now, since $0 < \bar{\lambda} < \gamma/2 < \bar{\rho}_2$ and $\bar{\omega} > 1$, we can deduce that $\bar{\alpha} + \bar{\tau} > 0$, which proves (3.22).

We next prove (3.23) by contradiction. Assume that $\bar{\tau}\bar{\omega} + \bar{\beta} \geq 0$; then from equality (3.21) we can write that

$$\bar{\lambda}^2 - \bar{\lambda}\bar{\rho}_2 = \bar{\lambda}(\bar{\lambda} - \bar{\rho}_2) \geq 0,$$

which implies, since $\bar{\lambda} > 0$, that $\bar{\lambda} \geq \bar{\rho}_2 \geq \gamma/2$ and contradicts the assumption $\bar{\lambda} < \gamma/2$.

We finally prove (3.24). Multiplying the left equality in (3.10) by $\bar{x}_1^T C$ gives

$$\bar{x}_1^T C^2 \bar{x}_1 + \bar{x}_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} \bar{x}_2 = (\bar{\alpha} + \bar{\tau})\|\bar{x}_1\|^2 = \bar{\lambda}\bar{x}_1^T C\bar{y}_1.$$

Combining this last equality with (3.22) and the Cauchy–Schwarz inequality, we obtain

$$0 < (\bar{\alpha} + \bar{\tau})\|\bar{x}_1\|^2 = \bar{\lambda}\bar{x}_1^T C\bar{y}_1 \leq \bar{\lambda}\|C\bar{x}_1\|\|\bar{y}_1\| = \bar{\lambda}\sqrt{\bar{\alpha}}\|\bar{x}_1\|\|\bar{y}_1\|, \tag{3.27}$$

where the last equality derives from the definition of $\bar{\alpha}$ in (3.18). Squaring both sides of (3.27), we also have that

$$(\bar{\alpha}+\bar{\tau})^2\|\bar{x}_1\|^2 \leq \bar{\lambda}^2\bar{\alpha}\|\bar{y}_1\|^2 \leq \bar{\lambda}^2\bar{\alpha}(\|\bar{y}_1\|^2+\|\bar{y}_2\|^2).$$

Combining this last inequality with (3.25) and dividing by $\|\bar{x}_2\|^2$ gives

$$(\bar{\alpha}+\bar{\tau})^2\bar{\omega} \leq (\bar{\alpha}^2+2\bar{\alpha}\bar{\tau})\bar{\omega}+\bar{\alpha}\bar{\beta},$$

which, after simplification, yields the desired result (3.24). ∎

3.3. Analyzing possible lower bounds. In the previous section, we have assumed the existence of a positive eigenvalue $\bar{\lambda}$ of $\bar{\mathcal{A}}$ satisfying $\bar{\lambda} < \gamma/2$ and shown, under the assumption that $c_{\min} > 0$, that $\bar{\lambda}$ and its associated eigenvector $\bar{x}$ satisfy the relations (3.20)–(3.24). In order to refine the positive lower bound in (1.4) as given by Rusten and Winther [10], we proceed in two steps by introducing two optimization problems successively, whose optimal solution will provide the desired refined positive lower bound on the eigenvalue $\bar{\lambda}$.

To build the feasible domain of the first of these two optimization problems, we relax the relations (3.20)–(3.24) by relaxing the quantities $\bar{\lambda}$, $\bar{\tau}$, and $\bar{\omega}$ in these relations, which now become the variables λ, τ, and ω verifying

$$\bar{\rho}_1 \leq \lambda \leq \frac{\bar{\rho}_2}{2} \quad \text{and} \quad 1 \leq \omega \leq \omega_{\max},$$

where $\omega_{\max}$ is an upper bound satisfying $\omega_{\max} \geq \bar{\omega}$. The constraints of this optimization problem, which we shall denote as $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$, are defined by the following relations:

$$\left.\begin{aligned}
&-\lambda^2+\lambda\bar{\rho}_1+\bar{\alpha}+\tau=0, && \text{(3.28a)}\\
&-\lambda^2+\lambda\bar{\rho}_2+\bar{\beta}+\tau\omega=0, && \text{(3.28b)}\\
&\bar{\rho}_1 \leq \lambda \leq \frac{\bar{\rho}_2}{2}, && \text{(3.28c)}\\
&1 \leq \omega \leq \omega_{\max}, && \text{(3.28d)}\\
&\tau \geq -\bar{\alpha}, && \text{(3.28e)}\\
&\tau\omega \leq -\bar{\beta}, && \text{(3.28f)}\\
&\tau^2\omega \leq \bar{\alpha}\bar{\beta} && \text{(3.28g)}
\end{aligned}\right\} \equiv \mathcal{F}(P).$$

We then minimize λ over the set (λ, τ, ω) satisfying these constraints; i.e., we consider the following optimization problem:

$$P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max}) = \min_{(\lambda,\tau,\omega)\in\mathcal{F}(P)} \lambda. \tag{3.29}$$

Note that $\mathcal{F}(P)$ is nonempty, since $(\bar{\lambda}, \bar{\tau}, \bar{\omega}) \in \mathcal{F}(P)$, and that λ no longer represents an eigenvalue of the matrix $\bar{\mathcal{A}}$ in this problem. Instead, the optimal value λ_0 of (3.29), whose existence is guaranteed by the compactness of the feasible set $\mathcal{F}(P)$ (see the Weierstrass theorem in [6]), gives a lower bound on $\bar{\lambda}$ (since $(\bar{\lambda}, \bar{\tau}, \bar{\omega}) \in \mathcal{F}(P)$). To assess the compactness of $\mathcal{F}(P)$, observe that τ in (3.29) satisfies $\tau \leq 0$ by (3.28f), since $\omega \geq 1$ and $\bar{\beta} \in [0, 1]$ by (3.18).

We shall now study the global optimum of the optimization problem (3.29). In that respect, we establish a first useful lower bound in the next lemma, whose proof is rather quick and given in Appendix D.

LEMMA 3.4. *Given the scalar values* $\bar{\alpha} > 0$, $\bar{\beta} \geq 0$, $\omega \geq 1$, *and* $\tau \leq 0$ *satisfying* (3.28g), *the system of equations* (3.28a) *and* (3.28b) *in* λ *has a unique positive solution satisfying*

$$\lambda \geq \frac{\omega\bar{\rho}_1 + \bar{\rho}_2 + \sqrt{\bar{\Delta}}}{2(\omega+1)}, \tag{3.30}$$

where $\bar{\Delta} = (\omega\bar{\rho}_1 + \bar{\rho}_2)^2 + 4(\omega+1)\big(\sqrt{\bar{\alpha}\omega} - \sqrt{\bar{\beta}}\big)^2$.

Our study of an optimal solution of the optimization problem (3.29) continues with the identification of the constraints which are potentially active at optimality.

THEOREM 3.5. *Consider problem* $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$ *defined in* (3.29) *where* $\omega_{\max} \geq \bar{\omega}$ *and* $\bar{\omega}, \bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}$, *and* $\bar{\beta}$ *are given by* (3.16)–(3.18). *Then the only constraints in* (3.29) *possibly active at a global solution are*

$$\omega \leq \omega_{\max} \qquad \textit{and} \qquad \tau^2\omega \leq \bar{\alpha}\bar{\beta}.$$

Proof. First, let us prove that the lower bound in constraint (3.28d) ($\omega = 1$) is not active. By Lemma 3.4 when $\omega = 1$, we have that

$$\lambda \geq \frac{\bar{\rho}_1 + \bar{\rho}_2 + \sqrt{(\bar{\rho}_1 + \bar{\rho}_2)^2 + 8(\sqrt{\bar{\alpha}} + \sqrt{\bar{\beta}})^2}}{4} \geq \frac{(\bar{\rho}_1 + \bar{\rho}_2)}{2} > \frac{\bar{\rho}_2}{2},$$

since $\bar{\rho}_1 > 0$. This is incompatible with (3.28c).

In a second step, we prove that the constraint (3.28e) is not active. Assuming that $\tau = -\bar{\alpha}$, (3.28a) and (3.28b) then become

$$-\lambda^2 + \lambda\bar{\rho}_1 = 0 \quad \text{and} \quad -\lambda^2 + \lambda\bar{\rho}_2 + \bar{\beta} - \bar{\alpha}\omega = 0, \tag{3.31}$$

respectively. The first equation in (3.31) implies that either $\lambda = 0$ or $\lambda = \bar{\rho}_1$. Since $\lambda \geq \bar{\rho}_1 > 0$ by (3.28c), we have that its unique solution is $\lambda = \bar{\rho}_1$, and it follows that the second equation in (3.31) becomes $-(\bar{\rho}_1)^2 + \bar{\rho}_1\bar{\rho}_2 + \bar{\beta} - \bar{\alpha}\omega = 0$, or equivalently,

$$\bar{\alpha}\omega = \bar{\rho}_1(\bar{\rho}_2 - \bar{\rho}_1) + \bar{\beta}. \tag{3.32}$$

Multiplying (3.32) by $\bar{\alpha}$, we obtain

$$\bar{\alpha}^2\omega = \tau^2\omega = \bar{\alpha}(\bar{\rho}_1(\bar{\rho}_2 - \bar{\rho}_1) + \bar{\beta}).$$

Since $\bar{\alpha} > 0$, $\bar{\rho}_1 > 0$, and $\bar{\rho}_2 - \bar{\rho}_1 > 0$, we can deduce that $\tau^2\omega > \bar{\alpha}\bar{\beta}$, which contradicts (3.28g).

We next prove that (3.28f) is not active by contradiction. Assuming that $\tau\omega + \bar{\beta} = 0$, (3.28b) becomes $-\lambda^2 + \lambda\bar{\rho}_2 = 0$, so that either $\lambda = 0$ or $\lambda = \bar{\rho}_2$, which is impossible by (3.28c).

We finally prove that both bounds of (3.28c) are inactive at a global solution. First, $\lambda = \bar{\rho}_1$ implies by (3.28a) that (3.28e) is active, which is impossible by the second step of the proof. If $\lambda = \bar{\rho}_2/2$, then it is not a global solution, since $(\bar{\lambda}, \bar{\tau}, \bar{\omega}) \in \mathcal{F}(P)$ and $\bar{\lambda} < \gamma/2 \leq \bar{\rho}_2/2 = \lambda$ provides a lower objective function value. ∎

The last step in the study of the optimization problem (3.29) is to show that either one or the other of the two constraints identified in Theorem 3.5 is actually active, giving then fundamental information to raise a lower bound of the global optimal

solution λ_0, which we recall is itself a lower bound of the assumed existing positive eigenvalue $0 < \bar{\lambda} < \gamma/2$ of $\bar{\mathcal{A}}$.

Let us first assume that $\omega = \omega_{\max}$ (i.e., the inequality constraint (3.28d) is active) at a global solution of problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$ given by (3.29). The lower bound in Lemma 3.4, with $\omega = \omega_{\max}$, gives us a first possible value for a lower bound on the optimal solution of problem (3.29):

$$\lambda_0 \geq \frac{1}{2}\left(\frac{\omega_{\max}\bar{\rho}_1 + \bar{\rho}_2 + \sqrt{\bar{\Delta}}}{\omega_{\max}+1}\right), \tag{3.33}$$

where $\bar{\Delta} = (\omega_{\max}\bar{\rho}_1 + \bar{\rho}_2)^2 + 4(\omega_{\max}+1)\left(\sqrt{\bar{\alpha}\omega_{\max}} - \sqrt{\bar{\beta}}\right)^2$. We can observe that this lower bound actually depends essentially on the choice for the value of $\omega_{\max} \geq \bar{\omega}$, the other scalar values $\bar{\rho}_1$, $\bar{\rho}_2$, $\bar{\alpha}$, $\bar{\beta}$ being fixed in the setting of problem (3.29). Defining now the quantities

$$\widehat{\rho} = \frac{\omega_{\max}\bar{\rho}_1 + \bar{\rho}_2}{\omega_{\max}+1} \quad \text{and} \quad \widehat{\alpha} = \left(\sqrt{\bar{\alpha}}\sqrt{\frac{\omega_{\max}}{\omega_{\max}+1}} - \sqrt{\frac{\bar{\beta}}{\omega_{\max}+1}}\right),$$

we can rewrite (3.33) as

$$\lambda_0 \geq \frac{1}{2}\left(\widehat{\rho} + \sqrt{\widehat{\rho}^2 + 4\widehat{\alpha}^2}\right). \tag{3.34}$$

This lower bound thus holds in the case where $\omega = \omega_{\max} \geq \bar{\omega}$ at an optimal solution of problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$. Note also, for further use, that since $\omega_{\max}$ can a priori be taken as large as we want, one has that

$$\lim_{\omega_{\max}\to\infty} \widehat{\rho} = \bar{\rho}_1 \quad \text{and} \quad \lim_{\omega_{\max}\to\infty} \widehat{\alpha} = \sqrt{\bar{\alpha}}. \tag{3.35}$$

We now study the case where $\omega < \omega_{\max}$ to raise a second possible lower bound value. The minimum of these two bounds will finally enable us to give a refined lower bound on the set of positive eigenvalues of $\bar{\mathcal{A}}$. If $\omega < \omega_{\max}$ at a global solution of problem (3.29), the first thing to point out is that it is then necessary to have $\tau^2\omega = \bar{\alpha}\bar{\beta}$, that is, the constraint (3.28g) must be active. To see that, let us consider the most general first-order necessary optimality conditions that must hold at the optimal solution of problem (3.29), given by the F. John theorem (see [6, Theorem 3.1]), which states that there exist t, u, v, $\zeta \in \mathbb{R}$ not all equal to zero, such that

$$\begin{bmatrix}1\\0\\0\end{bmatrix} t - \begin{bmatrix}-2\lambda+\bar{\rho}_1\\1\\0\end{bmatrix} u - \begin{bmatrix}-2\lambda+\bar{\rho}_2\\\omega\\\tau\end{bmatrix} v - \begin{bmatrix}0\\-2\tau\omega\\-\tau^2\end{bmatrix} \zeta = \begin{bmatrix}0\\0\\0\end{bmatrix}, \tag{3.36}$$

with $\zeta \geq 0$ and the complementarity condition

$$\zeta(\bar{\alpha}\bar{\beta} - \tau^2\omega) = 0. \tag{3.37}$$

Note first that $\tau \neq 0$. Indeed, otherwise $\bar{\beta} \leq 0$ by (3.28f), implying that $\bar{\beta} = 0$, since $\bar{\beta} \in [0, 1]$, which in turn would give, by (3.28b),

$$-\lambda^2 + \lambda\bar{\rho}_2 = \lambda(\bar{\rho}_2 - \lambda) = 0,$$

so that $\lambda = 0$ or $\lambda = \bar{\rho}_2$, in contradiction with (3.28c). We next have that $\zeta \neq 0$, since otherwise $\tau v = 0$ by the third equality in (3.36), and thus $v = 0$, which in turn implies $u = 0$ by the second equality in (3.36), followed by $t = 0$ by the first equality in (3.36). This is incompatible with the assumption that t, u, v, and ζ cannot be all equal to zero. The complementarity condition (3.37) then ensures that the constraint (3.28g) must be active, i.e., $\tau^2\omega = \bar{\alpha}\bar{\beta}$. Note that, since we know that $\tau < 0$ (we have seen before that $\tau \leq 0$, and just above that $\tau \neq 0$ under the current assumption, e.g., $\omega < \omega_{\max}$), we can then set

$$\tau = -\sqrt{\frac{\bar{\alpha}\bar{\beta}}{\omega}}. \tag{3.38}$$

To continue our study of possible solutions when $\omega < \omega_{\max}$, let us now denote by τ_0 and $\omega_0 < \omega_{\max}$ the associated quantities to such a solution λ_0 of problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$. From (3.38), we then have $\tau_0 = -\sqrt{\bar{\alpha}\bar{\beta}/\omega_0}$, and remembering that $\bar{\beta} \neq 0$, we can set

$$\delta_0 = \sqrt{\frac{\bar{\alpha}\omega_0}{\bar{\beta}}}.$$

Observing that $\tau_0 = -\bar{\alpha}/\delta_0 = -\delta_0\bar{\beta}/\omega_0$, we can then rewrite (3.28a) and (3.28b) as

$$-\lambda_0^2 + \lambda_0\bar{\rho}_1 + \bar{\alpha}\left(1 + \frac{\tau_0}{\bar{\alpha}}\right) = -\lambda_0^2 + \lambda_0\bar{\rho}_1 + \bar{\alpha}\left(1 - \frac{1}{\delta_0}\right) = 0 \tag{3.39}$$

and

$$-\lambda_0^2 + \lambda_0\bar{\rho}_2 + \bar{\beta}\left(1 + \frac{\tau_0\omega_0}{\bar{\beta}}\right) = -\lambda_0^2 + \lambda_0\bar{\rho}_2 + \bar{\beta}\,(1 - \delta_0) = 0. \tag{3.40}$$

We also have by (3.28e) (which is inactive at a solution) that $\tau_0 = -\bar{\alpha}/\delta_0 > -\bar{\alpha}$, so that $\delta_0 > 1$.

In order to conclude our search for a refined positive lower bound in (1.4), we next consider a second (and last) constrained optimization problem, also built in the same spirit, i.e., with the aim to derive this time a lower bound on the above solution λ_0 of the optimization problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$, under the assumption that the associated value ω_0 is strictly less than $\omega_{\max}$. To this end, we relax the quantities λ_0, δ_0, $\bar{\alpha}$, and $\bar{\beta}$ and consider the optimization problem

$$\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2) \equiv \min_{(\lambda,\delta,\alpha,\beta)\in\mathcal{F}(\widetilde{P})} \lambda, \tag{3.41}$$

where the feasible set $\mathcal{F}(\widetilde{P})$ is now defined by

$$\left.\begin{aligned}
&-\lambda^2 + \lambda\bar{\rho}_1 + \alpha\left(1 - \frac{1}{\delta}\right) = 0, && \text{(3.42a)}\\
&-\lambda^2 + \lambda\bar{\rho}_2 + \beta\,(1 - \delta) = 0, && \text{(3.42b)}\\
&\bar{\rho}_1 \leq \lambda \leq \frac{\bar{\rho}_2}{2}, && \text{(3.42c)}\\
&1 \leq \delta \leq \delta_{\max}, && \text{(3.42d)}\\
&c_{\min}^2 \leq \alpha \leq 1, && \text{(3.42e)}\\
&0 \leq \beta \leq 1 && \text{(3.42f)}
\end{aligned}\right\} \equiv \mathcal{F}(\widetilde{P}),$$

with $\delta_{\max}$ an upper bound satisfying $\delta_{\max} \geq \delta_0 > 1$ and with $c_{\min} < 1$ (from the necessary condition raised in Lemma 3.2). Note that $\mathcal{F}(\widetilde{P})$ is nonempty, since $(\lambda_0, \delta_0, \bar{\alpha}, \bar{\beta}) \in \mathcal{F}(\widetilde{P})$ by (3.39), (3.40), (3.28c) satisfied by λ_0 and by (3.18). Again, the compactness of the feasible set $\mathcal{F}(\widetilde{P})$ guarantees the existence of an optimal value $\lambda_{\inf}$ for problem $\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ with $\lambda_{\inf} \leq \lambda_0$.

The next theorem identifies the constraints of $\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ which are potentially active at optimality.

THEOREM 3.6. *Consider problem $\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ defined in* (3.41)*, where $\delta_{\max} \geq \delta_0$ and $\bar{\rho}_1$ and $\bar{\rho}_2$ are given by* (3.17)*. Then the constraints in* (3.41) *possibly active at a global solution are*

$$\delta \leq \delta_{\max}, \qquad c_{\min}^2 \leq \alpha \leq 1, \qquad \textit{and} \qquad \beta \leq 1.$$

Proof. Let us first show that the lower bound in (3.42d) ($\delta = 1$) is not active at optimality. Indeed, if $\delta = 1$, we get by (3.42b) that $-\lambda^2 + \lambda\bar{\rho}_2 = \lambda\,(\bar{\rho}_2 - \lambda) = 0$, so that $\lambda = 0$ or $\lambda = \bar{\rho}_2$, in contradiction with (3.42c). Using the same argument, we have that the lower bound in (3.42f) ($\beta = 0$) is not active.

It remains to prove that both bounds of (3.42c) are inactive at a global solution. First, $\lambda = \bar{\rho}_1$ implies by (3.42a) that $\alpha(1 - \frac{1}{\delta}) = 0$, which is impossible, since $\alpha > 0$ and $\delta \neq 1$. If $\lambda = \frac{\bar{\rho}_2}{2}$, then it is not a global solution, since $\lambda_0 \leq \bar{\lambda} < \frac{\gamma}{2} < \frac{\bar{\rho}_2}{2} = \lambda$ provides a lower objective function value and $(\lambda_0, \delta_0, \bar{\alpha}, \bar{\beta}) \in \mathcal{F}(\widetilde{P})$. ∎

Similarly to the way we proceeded for problem (3.29), we first consider the case where $\delta = \delta_{\max}$ (i.e., the upper bound of (3.42d) is active) at a global solution of problem $\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2)$. Equation (3.42a), together with $\delta = \delta_{\max}$, gives

$$-\lambda^2 + \lambda\bar{\rho}_1 + \alpha\left(1 - \frac{1}{\delta_{\max}}\right) = 0,$$

whose roots are

$$\lambda_{1,2} = \frac{1}{2}\left(\bar{\rho}_1 \pm \sqrt{\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right)}\right),$$

where $\bar{\rho}_1^2 + 4\alpha\big(1 - \frac{1}{\delta_{\max}}\big) > 0$ as $\alpha > 0$ and $\delta_{\max} \geq \delta_0 > 1$. Excluding the negative root, one then gets

$$\lambda_{\inf} = \frac{1}{2}\left(\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4\alpha\left(1 - \frac{1}{\delta_{\max}}\right)}\right), \tag{3.43}$$

which raises a first lower bound for the solution λ_0 of problem $P(\bar{\rho}_1, \bar{\rho}_2, \bar{\alpha}, \bar{\beta}, \omega_{\max})$, under the assumption that the associated value ω_0 is strictly less than $\omega_{\max}$.

We next, and finally, consider the case where $\delta < \delta_{\max}$, and we shall see that it is then necessary that both the lower bound in constraint (3.42e) and the upper bound $\beta = 1$ in (3.42f) actually be active. To see that, we use Theorem 3.6 and again the F. John theorem (see [6, Theorem 3.1]) to state that there exist t, u, v, ζ, η, $\varphi \in \mathbb{R}$ not all equal to zero such that

$$\begin{bmatrix}1\\0\\0\\0\end{bmatrix} t - \begin{bmatrix}-2\lambda + \bar{\rho}_1\\ \frac{\alpha}{\delta^2}\\ 1 - \frac{1}{\delta}\\ 0\end{bmatrix} u - \begin{bmatrix}-2\lambda + \bar{\rho}_2\\ -\beta\\ 0\\ 1 - \delta\end{bmatrix} v - \begin{bmatrix}0\\0\\1\\0\end{bmatrix}\zeta - \begin{bmatrix}0\\0\\-1\\0\end{bmatrix}\eta + \begin{bmatrix}0\\0\\0\\-1\end{bmatrix}\varphi = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \tag{3.44}$$

with $\zeta \geq 0$, $\eta \geq 0$, $\varphi \geq 0$ and with the associated complementarity conditions

$$\zeta(c_{\min}^2 - \alpha) = 0, \tag{3.45}$$

$$\eta(\alpha - 1) = 0, \tag{3.46}$$

$$\varphi(\beta - 1) = 0. \tag{3.47}$$

If $\varphi = 0$, then $v = 0$ by the last equality in (3.44), since $\delta > 1$, and consequently $u = 0$ by the second equality in (3.44) and since $\alpha/\delta^2 \neq 0$. The first and third equalities of (3.44) then imply $t = 0$ and $\zeta = \eta$, respectively. Since t, u, v, ζ, η, φ cannot all be equal to zero, then one must have $\zeta = \eta \neq 0$, which implies, by the complementarity conditions (3.45) and (3.46), that $\alpha = 1 = c_{\min}^2$, which is impossible from the necessary condition raised in Lemma 3.2. Assume now that $\varphi > 0$. Then $\beta = 1$ by (3.47), and the last equality in (3.44) yields $(1 - \delta)v = \varphi$. This, together with $\varphi > 0$ and $\delta > 1$, implies that $v < 0$. The second equality in (3.44) then gives

$$-\frac{\alpha}{\delta^2}u + v = 0,$$

implying that $u < 0$, since $\alpha > 0$. By the third equality in (3.44), we get

$$-\left(1 - \frac{1}{\delta}\right)u - \zeta + \eta = 0,$$

that is, $\zeta - \eta > 0$, since $\delta > 1$. As (3.45) and (3.46) with $c_{\min}^2 < 1$ imply that either ζ or η must be zero, the only possibility is to have $\eta = 0$ and $\zeta > 0$ (since otherwise $\zeta = 0$ and $\eta < 0$, in contradiction with the sign condition $\eta \geq 0$ on the multiplier η). We thus have $\alpha = c_{\min}^2$ by (3.45).

Rewriting $\widetilde{P}(\bar{\rho}_1, \bar{\rho}_2)$ with $\beta = 1$ and $\alpha = c_{\min}^2$ finally results in a nonlinear system of two equations in (λ, δ),

$$\begin{cases} -\lambda^2 + \lambda\bar{\rho}_1 + c_{\min}^2\left(1 - \dfrac{1}{\delta}\right) = 0, \\ -\lambda^2 + \lambda\bar{\rho}_2 + (1 - \delta) = 0, \end{cases} \tag{3.48}$$

in which we look for the unique solution such that $\bar{\rho}_1 \leq \lambda \leq \frac{\bar{\rho}_2}{2}$ and $\delta \geq 1$ and from which we can deduce the second possible lower bound on the positive eigenvalues of $\bar{\mathcal{A}}$ as stated in the next theorem. The fact that this solution is indeed unique will be seen in the course of the proof that follows.

THEOREM 3.7. *Consider the nonlinear system of equations defined in* (3.48)*, where* $\bar{\rho}_1$ *and* $\bar{\rho}_2$ *are given by* (3.17)*. Then the unique positive solution satisfies*

$$\lambda_{sol} \geq \frac{\bar{\rho}_1 + \frac{4}{5}c_{\min}^2\bar{\rho}_2}{1 + \frac{4}{5}c_{\min}^2}. \tag{3.49}$$

Proof. Multiplying the first equation in (3.48) by δ and the second by $c_{\min}^2$ and summing, we have

$$-\lambda^2(\delta + c_{\min}^2) + \lambda(\bar{\rho}_1\delta + \bar{\rho}_2 c_{\min}^2) = 0.$$

This last equation has a single strictly positive solution that we can express as a function of δ,

$$\lambda(\delta) = \frac{\bar{\rho}_1\delta + \bar{\rho}_2 c_{\min}^2}{\delta + c_{\min}^2}.$$

Observing that its derivative

$$\lambda'(\delta) = \frac{\bar{\rho}_1(\delta + c_{\min}^2) - (\bar{\rho}_1\delta + \bar{\rho}_2 c_{\min}^2)}{(\delta + c_{\min}^2)^2} = \frac{(\bar{\rho}_1 - \bar{\rho}_2)c_{\min}^2}{(\delta + c_{\min}^2)^2} < 0,$$

since $\bar{\rho}_1 < \bar{\rho}_2$, we have that $\lambda(\delta)$ is a strictly decreasing function. Also observe that, in order to have a solution, the second equation in (3.48) requires that $\bar{\rho}_2^2 + 4(1-\delta) \geq 0$, that is, the largest possible value for δ to get a solution is $(\bar{\rho}_2^2 + 4)/4$. We can thus conclude, since $\bar{\rho}_2 \leq 1$ by (3.17), that

$$\lambda_{\text{sol}} \geq \lambda\left(\frac{\bar{\rho}_2^2 + 4}{4}\right) \geq \lambda(5/4) = \frac{\bar{\rho}_1 + \frac{4}{5}c_{\min}^2\bar{\rho}_2}{1 + \frac{4}{5}c_{\min}^2},$$

which ends the proof. ∎

3.4. Collecting the various results. In the previous subsection, we have raised several possible lower bounds for the set of positive eigenvalues of matrix $\bar{\mathcal{A}}$ in (3.1). Before gathering these various bounds to formulate our main result, let us redraw the plot of the reasoning underlying the technical parts and developments conducted so far.

The rationale was the following, starting from two ground basis assumptions. The first one is that we consider the existence of some very small positive eigenvalue of $\bar{\mathcal{A}}$ in (3.1), $0 \leq \bar{\lambda} < \gamma/2$, for some given cut-off value $\gamma \in [\mu_{\min}, 1]$. The second one is that $c_{\min} > 0$, with $c_{\min}$ being the minimum value for the cosines of the principal angles between the subspace spanned by the constraint equations and the subspace spanned by the eigenvectors of the (1,1) block matrix A associated to the eigenvalues strictly less than γ. From these two assumptions, we could raise first in Lemma 3.2 some important necessary conditions. Out of these, it was then possible to introduce specific scalar quantities in prescribed intervals, given by (3.16)–(3.19), respectively, and to show in Theorem 3.3 some equalities and inequalities that they naturally verify. We then introduced the optimization problem defined in (3.29), which resumes in finding the minimum value of λ within a compact constraint set $\mathcal{F}(P)$ of equalities and inequalities, defined by (3.28a)–(3.28g) and directly derived from the relations raised in Theorem 3.3. The global minimum λ_0 of this problem is naturally a lower bound on the positive eigenvalue $\bar{\lambda}$, since $\bar{\lambda}$ belongs to $\mathcal{F}(P)$. Now, the structure of this optimization problem, due to the number of parameters involved, does not allow us to easily derive some analytical formulation for a global solution λ_0. We therefore studied instead the possible active constraints at an optimal solution, so as to be able to transform the problem by eliminating some dependent variables. With careful study of the first-order necessary optimality conditions given by the F. John theorem (see [6, Theorem 3.1]), and with some logic arguments based on the complementarity conditions that are provided within these necessary conditions, we could reduce the possible states to only two alternative cases (see Theorem 3.5 and the discussion that follows). The first of these two cases, which corresponds to saturating the constraint $\omega = \omega_{\max}$ in (3.28d), provides a first possible lower bound value (3.34) for the optimal solution λ_0. The second and alternative case is to saturate the inequality constraint (3.28g), and it was used to eliminate one variable and to consider a second but simpler optimization problem, defined in (3.41), where the compact feasible set of constraints $\mathcal{F}(\widetilde{P})$, defined by (3.42a)–(3.42f), contains the solution λ_0 currently considered. Therefore, a new global solution λ_{inf} of this last optimization problem must verify $\lambda_{\text{inf}} \leq \lambda_0$, and it provides a second possible

lower bound value. Similarly to the way we proceeded for problem (3.29), we again identified, from the first-order necessary optimality conditions, two alternative feasible cases at a solution, the first one raising the possible value (3.43) for $\lambda_{\inf}$, and the second one enabling us to finally reduce the problem to the algebraic set of two nonlinear equations (3.48), whose positive solution can be bounded below by (3.49).

Gathering these three results together, we can now formulate our main result.

THEOREM 3.8. *Assume that the matrix $\bar{\mathcal{A}}$ in* (3.1) *(in which $Q^TQ = I_m$, $\mu_{\min} > 0$, and $\mu_{\max} = 1$) has an eigenvalue $\bar{\lambda}$ satisfying $0 \leq \bar{\lambda} < \gamma/2$ (with $\gamma \in [\mu_{\min}, 1]$), and let $C \in \mathbb{R}^{p\times p}$ given by* (3.6) *be such that $c_{\min} = \min_{i=1:p}\{\cos\theta_i\} > 0$, together with $p < m$ and $m < n - p$. Then the eigenvalues of $\bar{\mathcal{A}}$ are bounded within*

$$\left[\frac{\mu_{\min} - \sqrt{\mu_{\min}^2 + 4}}{2}, \frac{1-\sqrt{5}}{2}\right] \bigcup \left[b_{\inf}, \frac{1+\sqrt{5}}{2}\right], \tag{3.50}$$

where $b_{\inf} = \min\left(\frac{\gamma}{2}, \frac{\mu_{\min} + \frac{4}{5}c_{\min}^2\gamma}{1+\frac{4}{5}c_{\min}^2}\right)$.

Proof. As proved below, the lower bound $b_{\inf}$ for the set of positive eigenvalues of $\bar{\mathcal{A}}$ results from the minimum of the three bounds obtained in (3.34), (3.43), and (3.49). From the rationale recalled just above, the minimum of these three values actually provides a lower bound on any existing positive eigenvalue of matrix $\bar{\mathcal{A}}$ that would be less than $\gamma/2$. This main assumption implies first that the minimal value for a lower bound that we are able to raise cannot be greater than $\gamma/2$. The second value included in the definition of $b_{\inf}$ is simply derived by replacing $\bar{\rho}_1$ and $\bar{\rho}_2$ in (3.49) by their minimal possible values $\mu_{\min}$ and γ, respectively.

Remember that the first two lower bounds in (3.34) and (3.43) have been obtained when considering that $\omega = \omega_{\max}$ in (3.28d) and $\delta = \delta_{\max}$ in (3.42d), respectively. As both values of $\omega_{\max}$ and $\delta_{\max}$ can a priori be taken as large as we want, we can notice that the two lower bounds (3.34) and (3.43) have a limit when $\omega_{\max} \to \infty$ and $\delta_{\max} \to \infty$, respectively, using (3.35) for the former. Taking into account that $\bar{\alpha} \geq c_{\min}^2$ for (3.34) and $\alpha \geq c_{\min}^2$ by (3.42e) for (3.43), both of these two limits can actually be bounded below by the same value,

$$\frac{1}{2}\left(\mu_{\min} + \sqrt{\mu_{\min}^2 + 4c_{\min}^2}\right). \tag{3.51}$$

At last, we can see that the value in (3.51) is strictly greater, for any choice of $0 < c_{\min} \leq 1$ and $\mu_{\min} \leq \gamma \leq 1$, than the second value raised in the definition of $b_{\inf}$ above. Indeed, proving this inequality is equivalent to showing that

$$\sqrt{\mu_{\min}^2 + 4c_{\min}^2} > \frac{5\mu_{\min} + 8c_{\min}^2\gamma - 4c_{\min}^2\mu_{\min}}{5 + 4c_{\min}^2},$$

and squaring both sides and subtracting, it is finally also equivalent to $64c_{\min}^6 + (160 - 64\gamma^2 + 64\gamma\mu_{\min})c_{\min}^4 + (100 + 40\mu_{\min}^2 - 80\gamma\mu_{\min})c_{\min}^2 > 0$. This last inequality is actually always true, since $c_{\min} > 0$, and since $(160 - 64\gamma^2 + 64\gamma\mu_{\min}) \geq 96$ and $(100 + 40\mu_{\min}^2 - 80\gamma\mu_{\min}) \geq 20$ (because $0 < \mu_{\min} \leq \gamma \leq 1$). Consequently, for sufficiently large values of $\omega_{\max}$ and $\delta_{\max}$, the three lower bounds given in (3.34), (3.43), and (3.49) can all be bounded below by the single value

$$\frac{\mu_{\min} + \frac{4}{5}c_{\min}^2\gamma}{1 + \frac{4}{5}c_{\min}^2}. \tag{3.52}$$

∎

We can easily see that the lower bound (3.52) belongs to the interval $]\mu_{\min}, \gamma[$ as a convex combination of $\mu_{\min}$ and γ, since $\mu_{\min} < \gamma$ and $c_{\min} > 0$. In that respect, the lower bound of the right interval in (1.4) is refined. We can also observe that, when $c_{\min} \to 0$, we have that $b_{\inf} \to \mu_{\min}$, which is consistent with the result given by Rusten and Winther [10]. At last, the bound gets close to $\mu_{\min}$ only when $c_{\min}^2 \simeq \mu_{\min}$.

4. Synthesis and perspectives. The analysis that has enabled us to draw the conclusions in Theorem 3.8 is actually independent of the splitting of the spectrum of the (1,1) block matrix A, which is defined by the choice of the cut-off value $\gamma \in [\mu_{\min}, 1]$ (with the maximum eigenvalue of A being equal to 1 after appropriate initial scaling). We can therefore vary the value of γ and consider the maximum of the values of $b_{\inf}$ in (3.50) as a function of γ, to propose a *maximal least value* for the positive eigenvalues of matrix $\bar{\mathcal{A}}$ in (3.1). The only issue, when varying γ, is to take care that the dimension p of the invariant subspace of A associated to all eigenvalues strictly less than γ verifies $p < \min(m, n-m)$, so as to ensure that the decomposition proposed in (3.6) remains valid. These last considerations lead to the conclusive result below.

THEOREM 4.1. *Consider the matrix $\bar{\mathcal{A}}$ in* (3.1)*, with $Q^TQ = I_m$, $\mu_{\min} > 0$, and $\mu_{\max} = 1$. Let $0 < \mu_{\min} = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_n = \mu_{\max} = 1$ be the eigenvalues of the symmetric and positive definite* (1,1) *block A ranged in increasing order. Denote by r the minimum dimension* $\min(m, n-m)$*, and consider the decreasing sequence*

$$c_1 \geq c_2 \geq \cdots \geq c_p \geq \cdots \geq c_r \geq 0,$$

where c_p is the minimum cosine value of the principal angles between $\mathcal{I}m(Q)$ and $\mathcal{I}m(U_p)$, with U_p being the matrix made with the p eigenvectors of A associated to the p smallest eigenvalues $\{\mu_1, \ldots, \mu_p\}$. Then the minimum positive eigenvalue of $\bar{\mathcal{A}}$ is bounded below by

$$\max_{1 \leq p \leq r-1} \min\left(\frac{\mu_{p+1}}{2}, \frac{\mu_{\min} + \frac{4}{5} c_p^2 \mu_{p+1}}{1 + \frac{4}{5} c_p^2} \right). \tag{4.1}$$

The proof of this theorem is straightforward, as a consequence of Theorem 3.8. We have just a few comments, however. We do not need to assume that the cosines are isolated from 0, simply because the value of

$$\frac{\mu_{\min} + \frac{4}{5} c_p^2 \mu_{p+1}}{1 + \frac{4}{5} c_p^2}$$

is equal to $\mu_{\min}$ whenever $c_p = 0$, and the result from Rusten and Winther actually shows that $\mu_{\min}$ is an absolute lower bound for the positive eigenvalues of matrix $\bar{\mathcal{A}}$ in (3.1). In the eventuality that $\mu_r/2 < \mu_{\min}$, the above theorem would then give a bound lower than that of Rusten and Winther, but this is not really an issue. Indeed, one can replace in (4.1) the value $\mu_{p+1}/2$ by $\max(\mu_{\min}, \mu_{p+1}/2)$, to incorporate the result from Rusten and Winther at any rate. Finally, note that the sequence of minimum cosine values is necessarily decreasing, since the invariant subspaces U_p are embedded one into the other with increasing values of p.

The *maximal least value* in Theorem 4.1 results from a compromise between the distribution of the eigenvalues of the (1,1) block A and the distribution of the sequence of cosines as well. We illustrate this compromise on the small KKT system already introduced in section 2.2. We scale the largest eigenvalue of A, and orthonormalize

the constraint matrix B, which is essentially equivalent to applying the block diagonal preconditioner (2.1) to the saddle point linear system (1.1).

Figure 4.1 shows the various elements to illustrate the compromise in Theorem 4.1. The (targeted) real value of the minimum positive eigenvalue of $\bar{\mathcal{A}}$ is indicated by the horizontal solid (blue) line, and the value of $\mu_{\min}$ is indicated by the horizontal dashed (black) line, respectively. The increasing curve with diamonds (in green) shows the increasing sequence of eigenvalues μ_p, and the decreasing curve with circles (in pink) shows the decreasing sequence of the cosines squared c_p^2 for $1 \leq p \leq r-1$ ($r = 150$ in this example). The sequence of values involved in the bound (4.1), from which results the above mentioned compromise, is displayed in the solid (red) curve for $1 \leq p \leq r-1$. The right plot corresponds to the case where some cosines were raised above the value 0.362, as discussed in section 2.2, and where MINRES would converge in linear mode (see Figure 2.2), and the left one corresponds to the original case with nonmodified cosines and with a more erratic convergence profile. We can observe that, in both cases, the *maximal least value* (4.1) reached, indicated by a star on the solid (red) curve, is slightly closer to the real value of the minimum positive eigenvalue of $\bar{\mathcal{A}}$ than is $\mu_{\min}$.

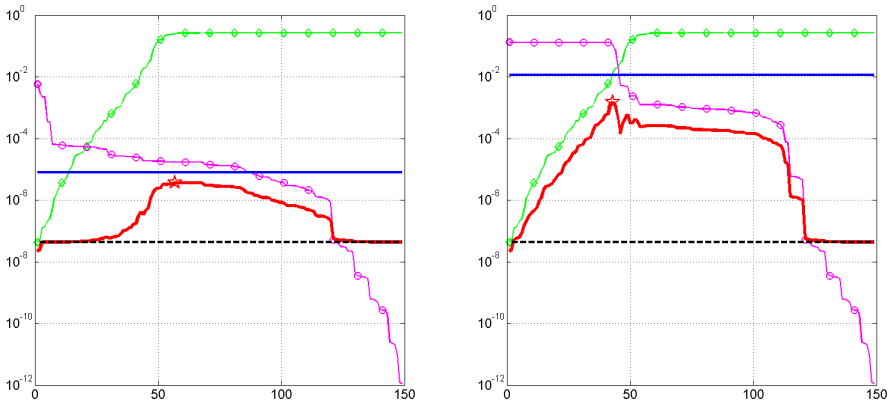

FIG. 4.1. *Illustration of the bound in* (4.1).

The theoretical results we presented allow us to clarify those situations where the spectral distribution of a saddle point matrix might effectively be affected by the presence of very small eigenvalues in its (1,1) block. The compromise between the distribution of the eigenvalues μ_p and the distribution of the sequence of cosines c_p that is involved is obviously problem dependent. Still, the knowledge of the value of the very first cosine can already be very informative, since the sequence of cosines is necessarily decreasing. Theorem 4.1 also indicates that there can be alternative directions to improve the condition number of a saddle point linear system, which is to either use a preconditioner that shifts the smallest eigenvalues of A to sufficiently larger values or to modify the system so as to improve the distribution of the cosines of the principal angles above. An idea could be, for instance, to augment the (1,1) block of $\bar{\mathcal{A}}$ in the spirit of what is done in [3]. At last, an erratic convergence profile of MINRES, after scaling of the (1,1) block and normalization (or near normalization) of the constraints, with a preconditioner of the type just above, for instance, may indicate a very rapidly decreasing sequence of cosines or a very small starting cosine value c_1.

It could be worth investigating the extensions of this theoretical analysis to the case of a semipositive definite (1,1) block A, using perturbation theory, for instance, and to the more general case where this (1,1) block is unscaled and the constraint equations are not necessarily orthonormalized as well. We still hope that this analysis can be useful to give some deeper insights in various situations.

Appendix A. Analytical study of example (1.5). To analyze a little further the case of the 3×3 matrix $\mathcal{A}_{cs}$ given in (1.5), consider its characteristic polynomial,

$$\chi(\lambda) = \lambda^3 - (1+\mu)\lambda^2 - (1-\mu)\lambda + (c^2 + s^2\mu),$$

which we rewrite as $\chi(\lambda) = \lambda^3 + Q(\lambda)$, where $Q(\lambda) = -(1+\mu)\lambda^2 - (1-\mu)\lambda + (c^2+s^2\mu)$. Observing that $\chi(\lambda) \geq Q(\lambda)$ for $\lambda \geq 0$ and that $\chi(0) = Q(0)$ and $\chi'(0) = Q'(0) = \mu - 1 < 0$, since $\mu < 1$, the two positive eigenvalues of $\mathcal{A}_{cs}$ are bounded below by the positive root of $Q(\lambda)$, that is,

$$\begin{aligned}
\lambda^+(\mathcal{A}_{cs}) &\geq \frac{(\mu-1) + \sqrt{(1-\mu)^2 + 4(1+\mu)(c^2+s^2\mu)}}{2(1+\mu)} \\
&= \frac{1}{2}\left(\sqrt{1 + 4\,\frac{c^2+s^2\mu^2}{(1+\mu)^2}} - \frac{1-\mu}{1+\mu}\right) \\
&= 2\,\frac{c^2(1-\mu)+\mu}{\sqrt{(1+\mu)^2 + 4(c^2(1-\mu^2)+\mu^2)} + 1 - \mu},
\end{aligned}$$

after simplifications, using $c^2 + s^2 = 1$ and multiplying both the top and bottom by the conjugate. This lower bound is $\mathcal{O}(\mu)$, as expected, when $c^2 = 0$, $\mathcal{O}(2\mu)$ when $c^2 = \mu$, and $\mathcal{O}\big(c^2/(1+c^2)\big)$, and thus bounded away from zero, when $\mu \ll c^2 \leq 1$.

Appendix B. Proof of Lemma 3.1. Multiplying the left equality in (3.9) by λx_1^T and using (3.11), we have

$$\lambda^2 \|x_1\|^2 = \lambda x_1^T M_\gamma x_1 + \lambda x_1^T C y_1 = \lambda \rho_1 \|x_1\|^2 + \lambda x_1^T C y_1. \tag{B.1}$$

Multiplying now the left equality in (3.10) by $x_1^T C$, we obtain

$$\lambda x_1^T C y_1 = x_1^T C^2 x_1 + x_1^T \begin{bmatrix} CS & 0 & 0 \end{bmatrix} x_2,$$

which together with (B.1) implies (3.12). Similarly, multiplying the right equality in (3.9) by λx_2^T and using (3.11), we get

$$\lambda^2 \|x_2\|^2 = \lambda \rho_2 \|x_2\|^2 + \lambda x_2^T \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} y_1 + \lambda x_2^T \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} y_2.$$

Then, using (3.10) to replace λy_1 and λy_2, we can derive

$$\lambda^2 \|x_2\|^2 = \lambda \rho_2 \|x_2\|^2 + x_2^T \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} \left(C x_1 + \begin{bmatrix} S & 0 & 0 \end{bmatrix} x_2\right) + x_2^T \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix} \begin{bmatrix} 0 & I & 0 \end{bmatrix} x_2,$$

implying (3.13). Finally, (3.14) and (3.15) immediately follow from (3.10).

Appendix C. Proof of Lemma 3.2. We first recall that $\bar{\mathcal{A}}$ is invertible, and consequently $\bar{\lambda} > 0$, under the hypotheses that are made (e.g., a KKT system with a

positive definite (1,1) block and a full rank constraint matrix). Let us show first that $\bar{x}_1$ is nonzero. Summing (3.12) and (3.13) and combining with (3.14) and (3.15), we have that

$$\bar{\lambda}^2(\|\bar{x}_1\|^2+\|\bar{x}_2\|^2)-\bar{\lambda}(\bar{\rho}_1\|\bar{x}_1\|^2+\bar{\rho}_2\|\bar{x}_2\|^2)=\bar{\lambda}^2(\|\bar{y}_1\|^2+\|\bar{y}_2\|^2), \tag{C.1}$$

with $\bar{\rho}_1\in[\mu_{\min},\mu_p]$, $\bar{\rho}_2\in[\mu_{p+1},1]$ and satisfying

$$\bar{x}_1^T M_\gamma \bar{x}_1=\bar{\rho}_1\|\bar{x}_1\|^2 \quad \text{and} \quad \bar{x}_2^T \widetilde{M}_\gamma \bar{x}_2=\bar{\rho}_2\|\bar{x}_2\|^2.$$

Observe that $\|\bar{x}_1\|^2+\|\bar{x}_2\|^2\neq 0$, since otherwise we would have that $\bar{x}_1=\bar{x}_2=0$ and consequently that $\bar{y}_1=\bar{y}_2=0$, since $\bar{\lambda}>0$, implying a zero eigenvector $\bar{x}$. From (C.1), we can then deduce, since $\bar{\lambda}>0$, that

$$\bar{\lambda}=\frac{\bar{\rho}_1\|\bar{x}_1\|^2+\bar{\rho}_2\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2+\|\bar{x}_2\|^2}+\bar{\lambda}\frac{\|\bar{y}_1\|^2+\|\bar{y}_2\|^2}{\|\bar{x}_1\|^2+\|\bar{x}_2\|^2}\geq\frac{\bar{\rho}_1\|\bar{x}_1\|^2+\bar{\rho}_2\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2+\|\bar{x}_2\|^2}=\bar{\rho}_1(1-\theta)+\bar{\rho}_2\theta,$$

where

$$\theta=\frac{\|\bar{x}_2\|^2}{\|\bar{x}_1\|^2+\|\bar{x}_2\|^2}\in[0,1].$$

Assuming $\|\bar{x}_2\|^2\geq\|\bar{x}_1\|^2$ (e.g., $\theta\geq 1/2$) leads to

$$\bar{\lambda}\geq\bar{\rho}_1(1-\theta)+\bar{\rho}_2\theta\in\left[\frac{\bar{\rho}_1+\bar{\rho}_2}{2},\bar{\rho}_2\right]$$

and in turn that $\bar{\lambda}\geq\gamma/2$ (since $\mu_{\min}\leq\bar{\rho}_1<\gamma\leq\mu_{p+1}\leq\bar{\rho}_2\leq 1$). This contradicts the assumption that $\bar{\lambda}<\gamma/2$, and hence it is necessary to have $\|\bar{x}_2\|^2<\|\bar{x}_1\|^2$, implying that $\bar{x}_1\neq 0$.

Assume now that $\bar{x}_2=0$; we then directly have, from the right equality in (3.9) and the left one in (3.10), that

$$S\bar{y}_1=0 \quad \text{and} \quad C\bar{x}_1=\bar{\lambda}\bar{y}_1. \tag{C.2}$$

Let us show that this also implies that

$$C\bar{x}_1=\bar{x}_1. \tag{C.3}$$

First, observe that from the assumption $c_{\min}>0$, one has that $\cos\theta_i>0$ for all $i=1,\dots,p$. If $\cos\theta_i=1$, then obviously (C.3) holds for index i. If $\cos\theta_i\neq 1$, implying that $\sin\theta_i\neq 0$ (since $\cos\theta_i^2+\sin\theta_i^2=1$), then the corresponding component in $\bar{y}_1$ is equal to zero by the first equality in (C.2), and so is the corresponding component in $\bar{x}_1$ by the second equality in (C.2), so that again (C.3) is satisfied for this index i. Finally, from (C.3) we get $C^2\bar{x}_1=C\bar{x}_1=\bar{x}_1$, and from (C.2) we also have $SC\bar{x}_1=\bar{\lambda}S\bar{y}_1=0$, so that we can rewrite (3.12) as

$$-\bar{\lambda}^2\|\bar{x}_1\|^2+\bar{\lambda}\bar{\rho}_1\|\bar{x}_1\|^2+\|\bar{x}_1\|^2=0,$$

and therefore, since $\bar{x}_1\neq 0$, we then get $-\bar{\lambda}^2+\bar{\lambda}\bar{\rho}_1+1=0$. The positive root of this

last equation in $\bar{\lambda}$ gives

$$\bar{\lambda} = \frac{\bar{\rho}_1 + \sqrt{\bar{\rho}_1^2 + 4}}{2} > 1,$$

i.e., $\bar{\lambda} > \mu_{\max} = 1$, which leads to a contradiction with the assumption that $\bar{\lambda} < \gamma/2$. Hence we must also have $\bar{x}_2 \neq 0$.

The necessary condition $c_{\min} < 1$ is also induced by the same considerations. Indeed, if $c_{\min} = 1$ (meaning that all cosines are equal to 1), then the first equalities in (C.2) and (C.3) both hold (since in this case $C = I$ and $S = 0$), and (3.12) leads again to $-\bar{\lambda}^2 + \bar{\lambda}\bar{\rho}_1 + 1 = 0$, with the same contradiction.

Appendix D. Proof of Lemma 3.4.

Proof. Multiplying (3.28a) by ω and adding it to (3.28b), we get the equation

$$(\omega + 1)\lambda^2 - (\omega\bar{\rho}_1 + \bar{\rho}_2)\lambda - \left(\omega(\bar{\alpha} + 2\tau) + \bar{\beta}\right) = 0,$$

whose roots are given by

$$\lambda_{1,2} = \frac{\omega\bar{\rho}_1 + \bar{\rho}_2 \pm \sqrt{\Delta}}{2(\omega + 1)},$$

where $\Delta = (\omega\bar{\rho}_1 + \bar{\rho}_2)^2 + 4(\omega + 1)(\omega\bar{\alpha} + 2\omega\tau + \bar{\beta})$. By (3.28g), together with $\omega > 0$, we have that $\tau^2\omega^2 \leq \bar{\alpha}\bar{\beta}\omega$, or equivalently that $\tau\omega \geq -\sqrt{\bar{\alpha}\bar{\beta}\omega}$, since $\tau \leq 0$, $\bar{\alpha} > 0$, and $\bar{\beta} \geq 0$. This last inequality implies that

$$\omega\bar{\alpha} + 2\omega\tau + \bar{\beta} \geq \omega\bar{\alpha} - 2\sqrt{\bar{\alpha}\bar{\beta}\omega} + \bar{\beta} = \left(\sqrt{\bar{\alpha}\omega} - \sqrt{\bar{\beta}}\right)^2 \geq 0,$$

so that, on the one hand, $\Delta \geq (\omega\bar{\rho}_1 + \bar{\rho}_2)^2$ for $\omega \geq 1$, yielding $\lambda_1 \leq 0$. On the other hand, it implies that $\Delta \geq \bar{\Delta}$, so that the unique positive solution of (3.28a) and (3.28b) is λ_2 and satisfies (3.30). ∎

REFERENCES

[1] M. Benzi, G. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[2] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Oxford University Press, Oxford, UK, 2005.

[3] G. H. Golub, C. Greif, and J. M. Varah, *An algebraic analysis of a block diagonal preconditioner for saddle point systems*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 779–792, https://doi.org/10.1137/04060679X.

[4] G. H. Golub and C. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.

[5] N. I. M. Gould and V. Simoncini, *Spectral analysis of saddle point matrices with indefinite leading blocks*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1152–1171, https://doi.org/10.1137/080733413.

[6] J. Hiriart-Urruty, *L'optimisation*, Presse Universitaire de France, Paris, 1996.

[7] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer Ser. Oper. Res. Financ. Eng., Springer-Verlag, Heidelberg, Berlin, New York, 2006.

[8] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405, https://doi.org/10.1137/0718026.

[9] I. Perugia and V. Simoncini, *Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.

[10] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904, https://doi.org/10.1137/0613054.

[11] M. Stoll and A. Wathen, *All-at-once solution of time-dependent Stokes control*, J. Comput. Phys., 232 (2013), pp. 498–515.

[12] C. Tannier, *Study of Block Diagonal Preconditioners Using Partial Spectral Information to Solve Linear Systems Arising in Constrained Optimization Problems*, Ph.D. thesis, Department of Mathematics, University of Namur, Namur, Belgium, 2016; available online from https://researchportal.unamur.be/en/studentTheses/study-of-block-diagonal-preconditioners-using-partial-spectral-in.